Combining heterogeneous inputs for the development of adaptive and multimodal interaction systems

David Griol, Jesús García-Herrero, José M. Molina

Applied Artificial Intelligence Group (GIAA), Computer Science Department, Carlos III University of Madrid, Spain. {david.griol,jesus.garcia,josemanuel.molina}@uc3m.es

KEYWORDS

Software agents
Multimodal fusion
Visual sensor networks
Surveillance applications
Spoken interaction
Conversational Agents
User Modeling
Dialog Management

ABSTRACT

In this paper we present a novel framework for the integration of visual sensor networks and speech-based interfaces. Our proposal follows the standard reference architecture in fusion systems (JDL), and combines different techniques related to Artificial Intelligence, Natural Language Processing and User Modeling to provide an enhanced interaction with their users. Firstly, the framework integrates a Cooperative Surveillance Multi-Agent System (CS-MAS), which includes several types of autonomous agents working in a coalition to track and make inferences on the positions of the targets. Secondly, enhanced conversational agents facilitate human-computer interaction by means of speech interaction. Thirdly, a statistical methodology allows modeling the user conversational behavior, which is learned from an initial corpus and improved with the knowledge acquired from the successive interactions. A technique is proposed to facilitate the multimodal fusion of these information sources and consider the result for the decision of the next system action.

1 Introduction

Research on multimodal interaction has grown considerably during the last decade as a consequence of the advent of innovative input interfaces, as well as the development of research fields such as speech interaction and natural language processing [GIBBON, D. *et al*. 2000],[LEMON, O. *et al*. 2012].

The widespread use of mobile technology implementing wireless communications such as smartphones and tablet-PCs enables a new type of advanced applications to access information. As the number of ubiquitous, connected devices continues to grow, the heterogeneity of client capabilities and the number of methods for accessing information services also increases. As a result, users can effectively access huge amounts of information and services from almost everywhere and through multimodal interaction.

Speech and natural language technologies allow users to communicate in a flexible and efficient manner, making possible to access applications in which traditional input interfaces cannot be used (e.g. in-car applications, access for disabled persons, etc). Also, speech-based interfaces work seamlessly with small devices and allow users to easily invoke local applications or access remote information. For this reason, multimodal conversational agents are becoming a strong alternative to traditional graphical interfaces which might not be appropriate for all users and/or applications [PIERACCINI, R., 2012], [LÓPEZ-CÓZAR, R. *et al*. 2005],

These systems go beyond both the desktop metaphor and the traditional speech-only interfaces by incorporating several communication modalities, such as speech, video, gaze, emotion recognition, gestures or facial expressions.

Multimodal conversational agents offer several additional advantages. Firstly, they can make

use of automatic recognition techniques to sense the environment allowing the user to employ different input modalities. Secondly, these systems typically employ several output modalities to interact with the user, which allows to stimulate several of his senses simultaneously, and thus enhance the understanding of the messages generated by the system.

In addition, the combination of modalities in the input and output allows to obtain more meaningful and reliable interpretations of the interaction context. This is, on the one hand, because complementary input modalities provide with non-redundant information which helps creating a richer model of the interaction.

On the other hand, redundant input modalities increase the accuracy and reduce the uncertainty of the information [CORRADINI, A. et al. 2003]. Finally, both the system and the user can choose the adequate interaction modalities to carry out the communication, thus enabling a better adaptation to environmental conditions such as light/acoustic conditions or privacy. Furthermore, the possibility to choose alternative ways of providing and receiving information allows disabled people to communicate with this type of system using the interaction modalities that best suits their needs.

Researchers have developed multimodal dialog systems for a number of applications, for example, interaction with mobile robots [LEVIN, E. et al. 2006] and information retrieval [HASEEL, L. *et al*. 2005]. These systems have also been applied to enhance the user-system interaction in homes [NAZARI, AA., 2005], [GAVER, WW. et al. 1992], academic centers [MARKOPOULOS, P. *et al*. 2005], hospitals [BRICON-SOUF, N. *et al*. 2007] and theme parks [NIGAY, L. *et al*. 1995].

Multimodality has been traditionally addressed from two perspectives. On the one hand, human-human multimodal communication. Within this area we can find in the literature studies concerned with speech-gesture systems [CATIZONE, R. *et al*. 2003], semiotics of gestures [RADFORD, L., 2003][FLECHA-GARCÍA, M.L., 2010], structure and functions of face-to-face communication [BAILLY, G. *et al*. 2010], emotional relations [COWIE, R. *et al*. 2003], [SCHULLER, S. *et al*. 2011]], and intercultural variations [ENDRASS, B. *et al*. 2011] [EDLUND, J. *et al*. 2008]. On the other hand, human-machine communication and interfaces. Topics of interest in this area include, among others, talking faces, embodied conversational agents [CASSELL, J. *et al*. 2000], integration of multimodal input, fission of multimodal output [WAHLSTER, W., 2003], and understanding of signals from speech, text, and visual images [BENESTY, J. *et al*. 2008].

In human conversation, speakers adapt their message and the way they convey it to their interlocutors and to the context in which the dialog takes place. The performance of a multimodal conversational agent also depends highly on its ability to adapt to the environmental conditions, such as other people speaking near the system or noise generated by other devices. This way, information related to the environment and users' presence and location is essential to achieve this adaptation [OSLAND, P., 2006], [LECH, T., 2005].

Adaptation can play a much more relevant role in speech-based applications [STRAUSS, P., 2010]. For example, users have diverse ways of communication. Novice users and experienced users may want the interface to behave completely differently, such as maintaining more guided versus more flexible dialogs. In these cases, processing context is not only useful to adapt the systems' behavior, but also to cope with the ambiguities derived from the use of natural language [SENEFF, S., 2007], [McCARTHY, J., 1987]. For instance, contextual information can be used to resolve anaphoric references depending on the context of the dialog or the user location.

In order to acquire this information, visual sensor networks (VSN) present a number of benefits. Firstly, the use of these networks is growing rapidly as powerful public safety and security tools (for instance, in airports [WEBER, M.E., 1994], sea environments [AVIS, P., 2003], railways or undergrounds [LO, B.P.2003], and other critical environments).

Secondly, the use of agents to develop VSNs provides additional advantages, like "reactivity" (agents can perceive and respond to a changing environment), "social ability" (by means of which agents interact with other agents), and "proactivity" (through which agents behave in a

goal-directed way). In addition, VSNs allow to know users current position (also considering users specific speeds, directions or even specific behaviors or physical features), but also to estimate users intentions and future actions (e.g., by detecting one or more users getting closer or moving away, looking at specific places, etc.).

In this work we present a novel architecture for the integration of visual sensor networks and speech-based interfaces. Our proposal is based on the multi-agent framework for deliberative camera-agents forming visual sensor networks described in [CASTANEDO, F., 2010]. In this framework, each camera is represented and managed by an individual software agent, called a surveillance-sensor agent [WOOLDRIDGE, M., 1995]. In addition, a visual fusion agent guarantees that objects of interest are successfully tracked across the whole area, assuring continuity and seamless transitions. The solution is a particular data fusion architecture integrating the data streams from different sensors and human voice to understand the situations.

As far as we are concerned, there are not previous works proposing the integration of the information provided by visual sensor networks to improve human machine interaction by means of conversational agents. To integrate speech interaction and visual sensor networks, we propose the incorporation of enhanced conversational agents [PIERACCINI, R., 2012], [LÓPEZ-CÓZAR, R., 2005]. This kind of agents can be defined as computer programs that accept natural language as input and produces natural language as output, engaging in a conversation with the user. To successfully manage the interaction with users, conversational agents usually carry out five main tasks: automatic speech recognition (ASR), natural language understanding (NLU), dialog management (DM), natural language generation (NLG), and text-to-speech synthesis (TTS). These tasks are usually implemented in different modules.

Each one of these tasks has its own characteristics and the selection of the most convenient model varies depending on certain factors: the goal of each module, or the capability of automatically obtaining models from training samples. The application of statistical approaches to dialog management has attracted increasing interest during the last decade [YOUNG, S., 2002]. Statistical models can be trained from real dialogs, modeling the variability in user behaviors. The final objective is to develop conversational agents that have a more robust behavior and are easier to adapt to different user profiles or tasks. The most extended methodology for machine-learning of dialog strategies consists of modeling human-computer interaction as an optimization problem using Partially Observable Markov Decision Processes (POMDPs) and reinforcement methods. However, they are limited to small-scale problems, since the state space would be huge and exact POMDP optimization would be intractable [WILLIAMS, J., *et al*. 2007].

In this paper we propose to incorporate two additional modules to generate enhanced conversational agents acting in conjunction with visual sensor networks. The first module, that we have called User Modeling Module, generates a prediction of the next user action by taking into account the previous interactions with the conversational agent. User profiles are considered in this module for a better prediction. The second module, that we have called Multimodal Fusion Module, generates the next input for the dialog manager by considering the spoken interaction and the information provided by the VSN.

The main contributions of this work are: (i) To provide a detailed architecture that considers heterogeneous information generated by cooperative surveillance multi-agent systems (CS-MAS) and conversational agents; (ii) To describe a multimodal fusion methodology that takes these information sources into account to generate and encode the input of the dialog manager in the conversational agent; (iii) To propose a statistical user modeling methodology to predict the current task of the dialog and the next user action; (iv) To provide a statistical methodology for dialog management that considers the data generated by the multimodal fusion and user modeling methodologies for the selection of the next system action.

After this brief introduction, the remaining of this paper is organized as follows. Section 2 describes the concept and main levels of data fusion from multiple sources. Section 3 presents the proposed framework for the integration of

visual sensor networks and speech-based interfaces. This section deals with the important challenges previously describes. Finally, Section 4 provides some conclusions and future works.

2 Data Fusion from Multiple Sources

Smart environments involve the deployment of a certain number of sensors (localization, cameras, microphones, etc) in a certain area to acquire data from the environment. Suitable procedures are needed to fuse data captured locally by sensors, in order to obtain an integrated view of the situation. In addition, the sensor networks and the associated procedures must be scalable, a difficult requirement when new heterogeneous sensors are expected to be integrated. The data fusion area studies problems arisen from the combination of several data sources. Fusion processes are classified according to the JDL (Joint Directors of Laboratories) model, the prevailing model to describe fusion systems [LIGGINS, M., 2009].

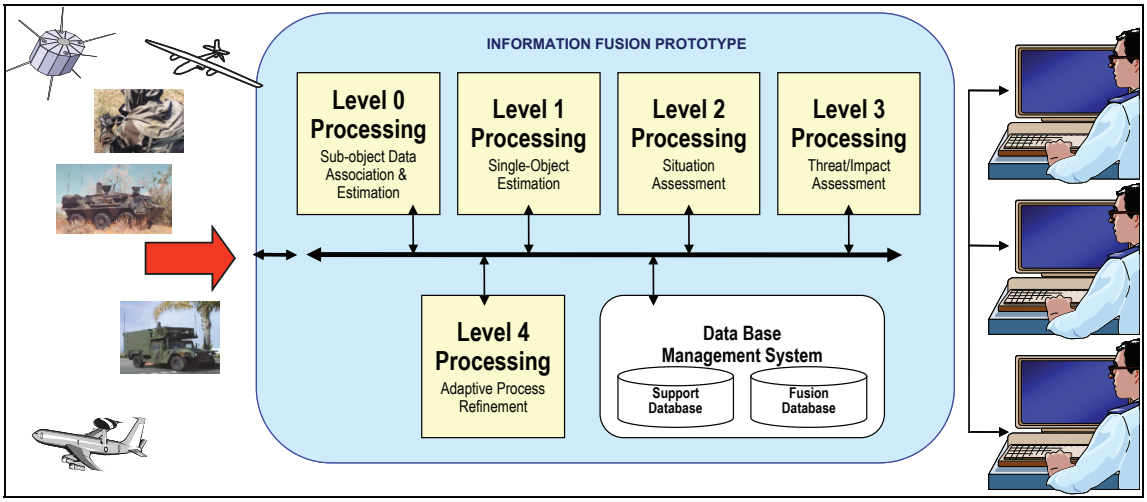

Fig. 1. The JDL Functional Model of the Information Fusion (adapted from [LIGGINS, M., 2009])

The processes, as Figure 1 shows, are classified according to the abstraction and the refinement of the involved entities. The canonical JDL model establishes five operational levels in the transformation of input signals to decision-ready knowledge, namely: signal feature assessment (L0), entity assessment (L1), situation assessment (L2), impact assessment (L3), and process assessment (L4).

Low-level data fusion, corresponding to JDL L0 and L1 levels, is the term used to designate procedures aimed to pre-process sensor signal and to estimate the properties of isolated objects. High-level information fusion procedures, corresponding to L2 and L3, aim to obtain a description of the relations between the objects in the perceived scenario. These relations are usually expressed with interpretable symbolic terms (e.g., actions, intentions, threats), instead of the usual numerical measures (e.g., density functions, movement vectors) calculated in L1. L4 tasks are aimed at planning and performing procedures to improve the whole fusion process, from low-level data acquisition to high-level situation assessment.

2.1 Level 0

The first decision is the physical installation of the sensors. The amount and the situation of sensors have a great influence on system cost and capabilities. For instance, it is convenient to arrange the cameras in a configuration that minimizes object occlusions and maximizes overlapping between fields of view, though this

is not always possible. The development of camera handover mechanisms to share information captured by a camera when an object moves to the field of view of an adjacent camera can be also considered in this step. Another essential step is calibration. Information must be aligned to a common reference frame.

Camera calibration, or common referencing, is the process to calculate the homography matrix that converts from the local coordinates of each camera to a global coordinate space. Calibration can be an off-line procedure (based on the correspondence of the position in the camera plane and in the global plane between of pre-defined landmarks) or an on-line procedure (based on the analysis of in-use system data; e.g., correspondences between automatically detected corners, edges, etc.).

2.2 Level 1

Object detection. There are various techniques for object detection, depending on sensors. For example, with video cameras, temporal differencing based on the calculation of the pixel-by-pixel difference between consecutive frames; background subtraction, optical flow, classification based on the identification of a pattern in the image with trained classifiers, etc. Object detection is not trivial, since in most cases the conditions of the watched environment change, and it is not possible to apply simple subtraction methods. Object tracking. Detected objects must be tracked over time; i.e., the system must segment the moving objects and assign consistent labels during their complete lifecycle. Object tracking is faced as a particular case of state estimation. It has been tackled by applying statistical prediction and inference methods, such as Kalman or particle filters, adapted to visual data association. These techniques are very sensitive to the particular conditions of the scenario, and therefore they may be insufficient in some applications. The incorporation of context knowledge has been regarded as essential to accomplish tracking requirements in complex scenarios with occlusions, illumination changes, and object deformations.

2.3 Level 2

Classification. Object identification and activity recognition are two fundamental classification tasks that must be performed in many applications based on sensor input. Object identification aims to determine the type of a tracked object; e.g., person, dish, box, etc. thus, it can be considered halfway between L1 and L2. This problem has been successfully faced by applying machine learning techniques to classify tracks according to extracted features: size, color, histogram, density, etc. Activity recognition, in turn, aims to identify that an activity is taking place. Two types of activities are distinguished: basic activities i.e. simple activities that cannot be decomposed into more simple actions (e.g., walking), and composite activities i.e., activities that are the result of various simple actions (e.g., laying the table). This problem remains unsolved in general applications, since it requires systems to develop cognitive capabilities close to human understanding. Recognition has been tackled with probabilistic methods (Markov models, Bayesian networks) and pattern recognition methods (neural networks, self-organization maps, k-means), though most approaches acknowledge the need of applying context knowledge to improve system performance.

2.4 Level 3

Situation assessment. Level 3 focuses on the estimation of the impact of a special situation to the application of interest. In other words, situation assessment is the process of detecting and evaluating particular situations that are of special relevance to the scenario because they relate to some type of threatening, critical situation, or any other special world state. This JDL level includes procedures aimed to the identification of abnormal and hazardous situations, which is especially relevant in some AmI domains; for example, Ambient Assisted Living applications require implementing proper mechanisms to react to an emergency situation if the user does not follow the normal sequence of activities, falls down, or abruptly interrupts an ongoing activity.

2.5 Level 4

Process enhancement. Process enhancement also known as fusion management is aimed at modifying the data acquisition and processing procedures to enhance results quality. Generally speaking, process enhancement consists in improving a fusion procedure by using feedback generated at a more abstract level. For instance, the behavior of a tracking algorithm can be changed once a general interpretation of the scene has been inferred; if the system recognizes that an object is moving out of the camera range through a door, the tracking procedure could be informed to be ready to delete this track in the near future.

3 Proposed Framework

The general architecture used for the development of multimodal applications can be separated in four different components: input modalities and their recognizers, output modalities and their respective synthesizers, the integration committee, and the application logic. Indeed, using multimodality efficiently implies a clear abstraction between the results of the user's input analysis, the processing of this input, answer generation and output modalities selection. As Figure 2 shows, this clear separation is achieved with help of the integration committee, responsible for management of all input and output modalities.

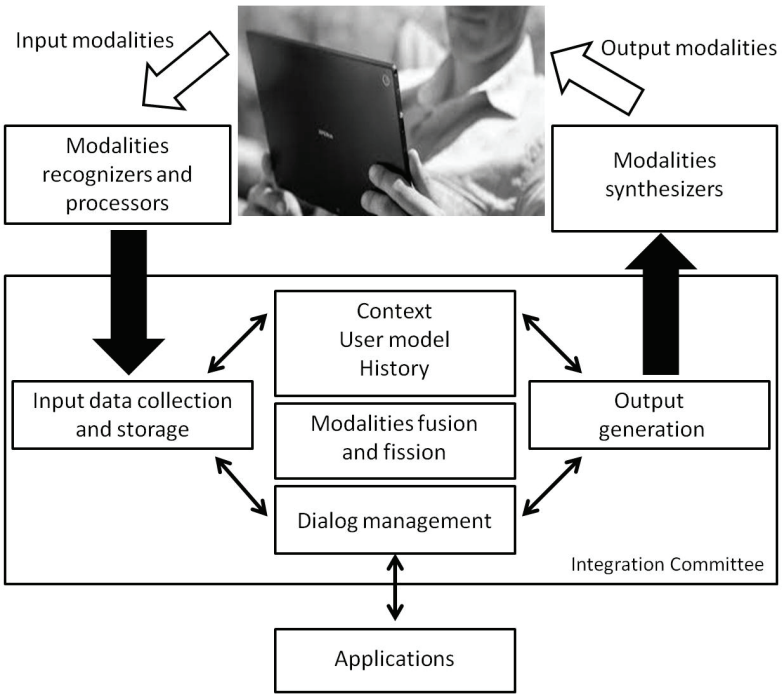

Fig. 2. General architecture for the generation of multimodal dialog systems

The integration committee can itself be separated in five different subcomponents. First, input modalities are collected into the input data collection and storage module, which is in charge of identifying and storing input data. The Modalities fusion and fission module manages input data prepares it for processing by the application logic. When the fusion and fission engines reach an interpretation, it is passed to the dialog management module.

In this paper, we propose a practical implementation of this general architecture for the construction of multimodal agents that allows speech interaction with their users and also makes use of specific techniques to consider visual information related to the context of the interaction. As described in the

introduction section, the proposed architecture to integrate visual sensor networks and speech interaction is based on [CASTANEDO, F., 2010].

As Figure 2 shows, different types of autonomous agents interact to fulfill this integration. The Surveillance-Sensor Agent tracks all the targets moving within its local field of view (FoV) and sends data to the Visual-Fusion Agent. It also sends information to the Context Agent. This agent is coordinated with other agents in order to improve surveillance quality. It can play different roles (individualized agent, object recognition agent, face recognition agent), each with different specific capabilities, but only one role at a time.

The Visual-Fusion Agent integrates the information sent from the associated surveillance-sensor agents. It analyzes the situation in order to manage the resources and coordinate the surveillance-sensor agents. This agent has the global view of the environment being monitored by all the surveillance-sensor agents.

It is in charge of creating the dynamic coalitions of surveillance-sensor agents using contextual information and the prediction of certain situations requiring a cooperative fusion process. This agent also integrates the information from the different cameras and assures continuity and seamless transitions.

The Recorder Agent belongs to a specific camera with recording features only [CASTANEDO, F., 2010]. The Planning Agent has a general vision of the whole scene. It makes inferences on the targets and the situation. The Context Agent provides monitored context-dependent information. This agent indicates the semantic distance between different surveillance-sensor agents. The context agent stores information about static objects that could provoke partial conclusions of the tracked targets but it also stores dynamic information about the scene [SÁNCHEZ, A.M., 2007]. The Interface Agent provides a graphical user interface that shows the evolution of the targets that are being tracked.

As described in [CASTANEDO, F., 2010], the coordination among Surveillance-Sensor Agents makes possible to jointly achieve a surveillance task. This way, the proposed CS-MAS architecture improves trajectory tracking by fusing data from several neighboring surveillance-sensor agents (camera agents in a visual sensor network), which are in a coalition.

In this paper, we propose the use of the information provided by the visual sensor network to facilitate the interaction with users by means of enhanced Conversational Agents.

As previously described, conversational agents integrate five main tasks: automatic speech recognition (ASR), natural language understanding (NLU), dialog management (DM), natural language generation (NLG), and text-to-speech synthesis (TTS).

Speech recognition is the process of obtaining the text string corresponding to an acoustic input [TSILFIDIS, A. *et al*. 2013] [LÓPEZ-CÓZAR, R. *et al*. 2008]. It is a very complex task as there is much variability in the input characteristics, which can differ depending on the linguistics of the utterance, the speaker, the interaction context and the transmission channel. Linguistic variability involves differences in phonetic, syntactic and semantic components that affect the voice signal. Inter-speaker variability refers to the big difference between speakers regarding their speaking style, voice, age, sex or nationality.

Once the conversational agent has recognized what the user uttered, it is necessary to understand what he said. Natural language processing is the process of obtaining the semantic of a text string [WU, W.-L. *et al*. 2010] [MINKER, W., 1999]. It generally involves morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge. Lexical and morphological knowledge allow dividing the words in their constituents distinguishing lexemes and morphemes. Syntactic analysis yields a hierarchical structure of the sentences, while semantic analysis extracts the meaning of a complex syntactic structure from the meaning of its constituents. In the pragmatic and discourse processing stage, the sentences are interpreted in the context of the whole dialog.

There is not a universally agreed upon definition of the tasks that a dialog manager has to carry. Traum and Larsson [TRAUM, D. *et al*. 2003] state that dialog managing involves four main tasks: i) updating the dialog context, ii) providing a context for interpretations, iii) coordinating other modules and iv) deciding the

information to convey and when to do it. Thus, the dialog manager has to deal with different sources of information such as the NLU results, database queries results, application domain knowledge, and knowledge about the users and the previous dialog history [GRIOL, D. *et al*. 2008].

Natural language generation is the process of obtaining texts in natural language from a non-linguistic representation. The simplest approach consists in using predefined text messages (e.g. error messages and warnings). Finally, a text-to-speech synthesizer is used to generate the voice signal that will be transmitted to the user.

As Figure 3 shows, two main modules have been incorporated to enrich the general architecture of a conversational agent previously described. As stated in the previous section, the User Modeling module considers the previous dialog interactions and specific users features (defined by means of user profiles) to calculate a prediction of the next user action. The Multimodal Fusion module takes as input this prediction, the current user utterance, and the information provided by the surveillance sub-system. Using this information this module generates the input of the dialog manager, which selects the next system action. The following subsections describe the statistical methodologies proposed for the development of these modules.

3.1 The user modeling module

Research in techniques for user modeling has a long history within the fields of language processing and speech technologies [SCHATZMANN, J., 2006]. The main purpose of a user intention model in this field is to improve the usability of a conversational agent through the generation of corpora of interactions between the system and the user model [GRIOL, D., 2011].

There are different levels in which the system can adapt to the user [JOKINEN, K., 2003]. The simplest one is through personal profiles in which the users have static choices to customize the interaction (e.g. whether they prefer a male or female system's voice), which can be further improved by classifying users into preference groups. Systems can also adapt to the users' environment, for example, Ambient Intelligence (AmI) applications such as ubiquitous proactive systems. The main research topics are the adaptation of systems to different expertise levels [HASEEL, L. *et al*. 2005], knowledge [FORBES-RILEY, K. M., 2004], and special needs of users. The latter topic is receiving a lot of attention nowadays in terms of how to make systems usable by handicapped and elderly people [HEIM, J. *et al*. 2005] [BATLINER, A. *et al*. 2004] [LANGNER, B. *et al*. 2005], and how to adapt them to user features such as age, proficiency in the interaction language [RAUX, A. *et al*. 2003] or expertise in using the system [HASEEL, L. *et al*. 2005].

Despite their complexity, these characteristics for the design of user centered multimodal interfaces are to some extent rather static, i.e. they are usually gathered a priori and not during the dialog, and thus they are not used to dynamically adapt the multimodal interface at some stage in the interaction. There is another degree of adaptation in which the system not only adapts to the messages conveyed during the interaction, but also to the user's intentions and emotional states [MARTINOVSKI, B. *et al*. 2003] [PRENDINGER, H. *et al*. 2003]. It has been demonstrated that many breakdowns in man-machine communication could be avoided if the machine was able to recognize the emotional state of the user and responded to it more sensitively, for instance, by providing more explicit feedback if the user is frustrated. Emotional intelligence not only includes the ability to recognize the user's emotional state, but also the ability to act on it appropriately [SALOVEY, P. *et al*. 1990].

Our proposed technique for user modeling simulates the user intention level by means of providing the next user dialog act in the same representation defined for the natural language understanding module. The lexical, syntactic and semantic information (e.g., words, part of speech tags, predicate-arguments structures, and name entities) associated to speaker *u*'s ith clause is denoted as c_i^u.

Our model is based on the proposed in [BANGALORE, S., 2008]. In this model, each user clause is modeled as a realization of a user action defined by a subtask to which the clause contributes, the dialog act of the clause, and the named entities of the clause. For speaker

u, DA_i^u denotes the dialog label of the i-th clause, and ST_i^u denotes the subtask label to which the i-th clause contributes. The dialog act of the clause is determined from the information about the clause and the previous dialog context (i.e., k previous utterances) as shown in Equation 1.

$$DA_i^u = \\ = \underset{d^u \in \mathcal{D}}{\operatorname{argmax}} P(d^u | c_i^u, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \quad (1)$$

In a second stage, the subtask of the clause is determined from the lexical information about the clause, the dialog act assigned to the clause according to Equation 1, and the dialog context, as shown in Equation 2.

$$ST_i^u = \\ = \underset{s^u \in \mathcal{S}}{\operatorname{argmax}} P(s^u | DA_i^u, c_i^u, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \quad (2)$$

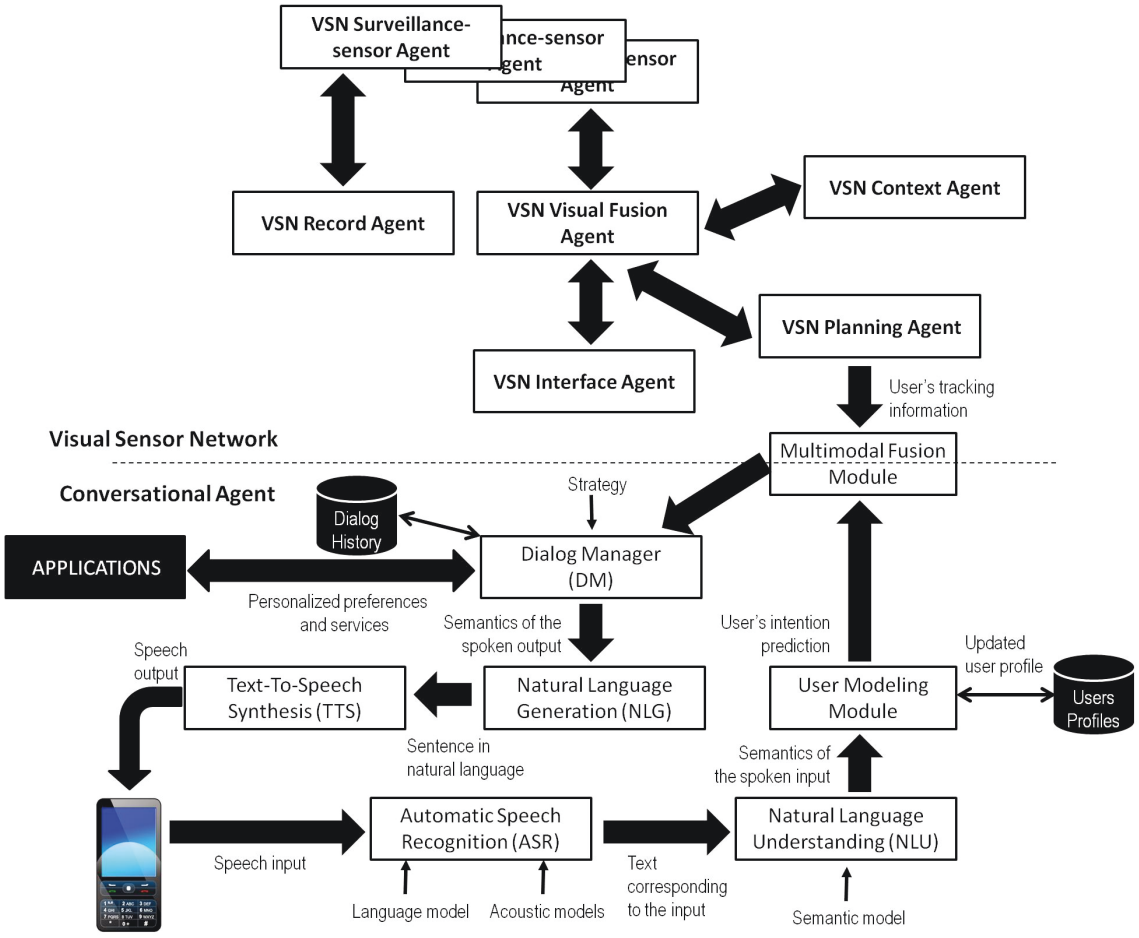

Fig. 3. Proposed multi-agent architecture to combine visual sensor networks and spoken interaction

In our proposal, we consider both static and dynamic features to estimate the conditional distributions shown in Equations 1 and 2. Dynamic features include the dialog act of each utterance and the task/subtask of each utterance. Static features include the words in each utterance (unigrams, bigrams, and trigrams), the part of speech tags in each utterance (unigrams,

bigrams, and trigrams), supertags in each utterance (unigrams, bigrams, and trigrams), and a set of features that has been included in a user profile. This profile is comprised of user's:
Id, which he can use to log in to the system;

- Gender;
- Experience, which can be either 0 for novel users (first time the user calls the system) or the number of times the user has interacted with the system;
- Skill level, estimated taking into account the level of expertise, the duration of their previous dialogs and the time that was necessary to access a specific content and the date of the last interaction with the system. A low, medium, high or expert level is assigned using these measures;
- Most frequent objective of the user;
- Reference to the location of the previous interactions and the corresponding objective and subjective parameters for the user.

3.2 Multimodal Fusion and Dialog Management

When dealing with multiple input sources, fusion of these input sources is a necessary feature of multimodal interaction creation tools. In fact, fusion of input data can be considered as one of the distinguishing features of multimodal interaction. Typical algorithms for decision-level fusion are frame-based fusion, unification-based fusion, and hybrid symbolic/statistical fusion [LALANNE, D., 2009]. Symbolic/statistical fusion [WU, L., 2002] is an evolution of standard symbolic unification-based approaches, which adds statistical processing techniques to the fusion techniques previously described. These kinds of hybrid fusion techniques have been demonstrated to achieve robust and reliable results.
The methodology that we propose to develop the multimodal fusion module considers the set of information sources (spoken interaction, user modeling, and video tracking) by using different machine-learning techniques. The main objective of this module is to successfully associate the visual situation detected by the VSN and the user interaction with the conversational agent.

As described in [BANGALORE, S., 2008], the conditional distributions shown in Equations 1 and 2 can be estimated by means of the general technique of choosing MaxEnt distribution that properly estimates the average of each feature in the training data [BERGER, A., 1996]. This can be written as a Gibbs distribution parameterized with weights as Equation 3 shows, where V is the size of the label set, X denotes the distribution of dialog acts or subtasks (DA_i^u or ST_i^u) and Φ denotes the vector of described features for user modeling

$$P(X = st_i|\phi) = \frac{e^{\lambda_{st_i}\cdot\phi}}{\sum_{st=1}^{V} e^{\lambda_{st_i}\cdot\phi}} \quad (3)$$

Each of the classes can be encoded as a bit vector such that, in the vector for class, the i-th bit is one and all other bits are zero. Then, one-versus-other binary classifiers are used as Equation 4 shows.

$$P(y|\phi) = 1 - P(\overline{y}|\phi) = \\ = \frac{e^{\lambda_y\cdot\phi}}{e^{\lambda_y\cdot\phi} + e^{\lambda_{\overline{y}}\cdot\phi}} = \frac{1}{1 + e^{-\lambda'_{\overline{y}}\cdot\phi}} \quad (4)$$

Where $\lambda_{\bar{y}}$ is the parameter vector for the anti-label $\bar{y}$ and $\lambda'_{\bar{y}} = \lambda_y - \lambda_{\bar{y}}$.
Once the users action prediction has been calculated, a prediction of the system action can also been generated using a similar process. Each system action is also defined in terms of the subtask to which it contributes and the dialog act to be performed. The determination of the system action, therefore, also proceeds in two stages: prediction of the system subtask (Equation 5) and prediction of the dialog act (Equation 6).

$$ST_i^a = \\ = \underset{s^a \in \mathcal{S}}{\operatorname{argmax}} P(s^u|ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \quad (5)$$

$$DA_i^a = \underset{d^a \in \mathcal{D}}{\operatorname{argmax}} P(d^a | ST_i^a, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \quad (6)$$

The dialog manager decides the next action of the conversational agent. In addition, it updates the dialog history, provides a context for interpreting the sentences, and coordinates the other modules of the multimodal system. Thus, the dialog manager has to deal with different sources of information such as the semantic interpretations of the users' utterances, database queries results, application domain knowledge, and knowledge about the users and the dialog history.

A conventional dialog manager maintains a state n such as a form or frame and relies on two functions for control, G and F. For a given dialog state n, G(n) = a decides which system action to output, and then after observation o has been received, F(n; o) = n_0 decides how to update the dialog state n to yield n_0. This process repeats until the dialog ends.

In a statistical approach, the conventional dialog manager is extended in three respects: firstly, its action selection function G(n) = a is changed to output a set of one or more (M) allowable actions given a dialog state n, G(n) = {a_1, a_2, …,a_M}. Next, its transition function F(n; o) = n_0 is extended to allow for different transitions depending on which of these actions was taken, F(n; a; o) = n_0.

In order to control the interactions with the user, our proposed statistical dialog management technique represents dialogs as a sequence of pairs (A_i, U_i), where A_i is the output of the dialog system (the system answer) at time *i*, and Ui is the semantic representation of the user turn (the result of the understanding process of the user input) at time *i*; both expressed in terms of dialog acts [GRIOL, D., 2008].

This way, each dialog is represented by:

$$(A_1, U_1), \cdots, (A_i, U_i), \cdots, (A_n, U_n)$$

where $A1$ is the greeting turn of the system, and U_n is the last user turn. We refer to a pair (A_i;U_i) as *Si*, the state of the dialog sequence at time *i*.

In this framework, we consider that, at time *i*, the objective of the dialog manager is to find the best system answer A_i. This selection is a local process for each time *i* and takes into account the previous history of the dialog, that is to say, the sequence of states of the dialog preceding time *i*:

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i | S_1, \cdots, S_{i-1}) \quad (7)$$

where set *A* contains all the possible system answers.

Following Equation 7, the dialog manager selects the following system prompt by taking into account the sequence of previous pairs (A_i, U_i). The main problem to resolve this equation is regarding the number of possible sequences of states, which is usually very large. To solve the problem, we define a data structure in order to establish a partition in this space, i.e., in the history of the dialog preceding time *i*). This data structure, which we call Interaction Register (IR), contains the following information:

- sequence of user dialog acts provided by the user throughout the previous
- history of the dialog (i.e., the output of the NLU module);
- predicted user dialog act (generated by means of Equation 1);
- predicted user subtask (generated by means of Equation 2);
- predicted user position (provided by the agents in the virtual sensor network
- as explained in [CASTANEDO, F., 2010]);
- predicted system dialog act (generated by means of Equation 5);
- predicted system subtask (generated by means of Equation 6);

After applying these considerations and establishing the equivalence relation in the histories of dialogs, the selection of the best Ai is given by Equation 8.

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i | IR_{i-1}, S_{i-1}) \quad (8)$$

We propose the use of a classification process to decide the next system action following the previous equation. Specifically, we propose a multilayer perceptron (MLP) for the classification, where the input layer receives the

current state of the dialog, which is represented by the term (IR_{i-1},A_i). The values of the output layer can be viewed as the a posteriori probability of selecting the different user intention given the current situation of the dialog. Figure 4 summarizes the operation of the proposed multimodal fusion and dialog management methodologies. As it can be observed, the user modeling module provides predictions of the next user dialog act and the current subtask of the dialog. Then, the system prediction module considers this information to generate the corresponding estimations for the system. The complete set of predicted values and the user position prediction provided by the planning agent are inputs of the fusion module to generate the interaction register. The dialog manager considers this register and the current user turn for the selection of the next system action.

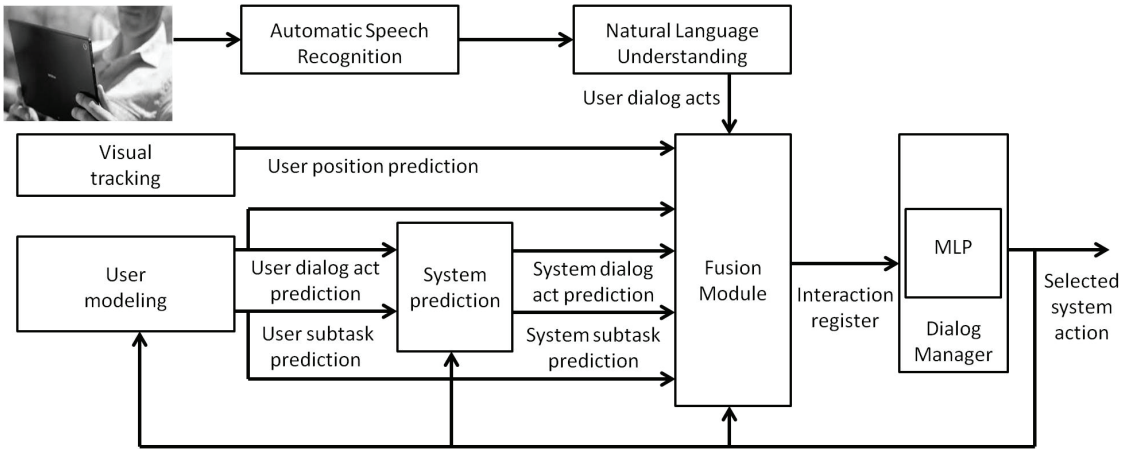

Fig. 4. Proposed multimodal fusion and dialog management methodologies for the development of conversational agents

4 Conclusions

In this paper we have described a framework to develop multi-agent systems that considers the information generated by cooperative surveillance systems to provide user-adapted spoken interaction. To do this, we propose the integration of enhanced conversational agents in the CS-MAS architecture described in [CASTANEDO, F., 2010].

Two main modules have been incorporated in the classical architecture of a conversational agent to achieve the integration between visual sensor networks and conversational agents. These modules respectively allow to predict the next user response for the conversational agent and carry out the fusion of visual and spoken information. The proposed multimodal fusion and dialog management techniques allow considering these heterogeneous information sources to select the next system action according to the current dialog and visual situations.

Although the different methodologies proposed to develop the described modules have been evaluated in previous works [CASTANEDO, F., 2010], [GRIOL, D., 2012], [BANGALORE, S., 2008], as a future work we propose the application of the described architecture to develop and evaluate a practical system in a real environment.

Acknowledgments

This work was supported in part by Projects MEyC TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485).

5 References

[AVIS, P., 2003] Avis, P., *Surveillance and Canadian maritime domestic security*, Canadian Military Journal, vol. 1, no. 4, pp. 9-15, 2003.

[BAILLY, G. *et al*. 2010] Bailly, G., Raidt, S., Elisei, F. *Gaze, conversational agents and face-to-face communication*. Speech Communication, 52(6), 598-612, 2010.

[BANGALORE, S. *et al*. 2008] Bangalore, S., G. D. Fabbrizio, and A. Stent. *Learning the Structure of Task-Driven Human-Human Dialogs*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 7, 1249-1259, 2008.

[BAKER, J. *et al*. 2008] Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C., Morgan, N., O'Shaughnessy, D. *Developments and directions in speech recognition and understanding*. IEEE Signal Processing Magazine 26(3), 75-80, 2009.

[BATLINER,A. *et al*. 2004] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M., Wong, M. *Towards multilingual speech recognition using data driven source/target acoustical units association*. Proc. of ICASSP'04. Montreal, Quebec, Canada, 521-524, 2004.

[BENESTY, J. *et al*. 2008] Benesty, J., Sondhi, M.M., Huang, Y. *Springer Handbook of Speech Processing*. Springer. 2008.

[BERGER, A. *et al*. 1996] Berger, A., S. Pietra, and V. Pietra. *A maximum entropy approach to natural language processing*, Comput. Linguist, 22(1), 39-71, 1996.

[BOHUS, D. *et al*. 1996] Bohus, D., Rudnicky, A. *RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda*. In: Proc. of 8th European Conference on Speech Communication and Technology (Eurospeech'03), pp. 597-600. Geneva, Switzerland, 2003.

[BRICON-SOUF, N. *et al*. 2007] Bricon-Souf N, Newman CR. *Context awareness in health care: A review*. International journal of medical informatics 76, 2-12, 2007.

[CASSELL, J. *et al*. 2000] Cassell, J., Sullivan, J., Prevost, S., Churchill, E.F. *Embodied Conversational Agents*. The MIT Press, 2000.

[CASTANEDO, F. *et al*. 2010] Castanedo, F., J. García, M. A. Patricio, and J. M. Molina. *Data fusion to improve trajectory tracking in a Cooperative Surveillance Multi-Agent Architecture*, Information Fusion, vol. 11, 243-255, 2010.

[CATIZONE, R. *et al*. 2003] Catizone, R., Setzer, A., Wilks, Y. *Multimodal Dialogue Management in the COMIC Project*. In: Proc. of EACL'03 Workshop on Dialogue Systems: interaction, adaptation, and styles of management. Budapest, Hungary, 25-34, 2003.

[CORRADINI, A. *et al*. 2003] Corradini A, Mehta M, Bernsen N, Martin J, Abrilian S. *Multimodal input fusion in human-computer interaction*. In: Proc. of the NATO-ASI Conference on Data Fusion for SituationMonitoring, IncidentDetection, Alert and Response Management, Yerevan, Armenia, 2003.

[COWIE, R. *et al*. 2003] Cowie, R., Cornelius, R. *Describing the emotional states that are expressed in speech*. Speech Communication, 40(1-2), 5-32, 2003.

[EDLUND, J. *et al*. 2008] Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson A. *Towards human-like spoken dialogue systems*. Speech Communication, 50 (8-9), 630-645, 2008.

[ENDRASS, B. *et al*. 2011] Endrass, B., Rehm, M., André, E. *Planning Small Talk behavior with cultural influences for multiagent systems*. Computer Speech & Language, 25(2), 158-174, 2011.

[FLECHA-GARCÍA, M.L., 2010] Flecha-García, M.L. *Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English*. Speech Communication, 52(6), 542-554, 2010.

[FORBES-RILEY, K. M., 2004] Forbes-Riley, K. M., Litman, D. *Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters*. In: Proc. of HLT-NAACL'04, New York, USA, 264-271, 2004.

[GAVER, WW. *et al*. 1992] Gaver WW. *Using and creating auditory icons*. SFI studies in the sciences of complexity, Addison Wesley Longman, 1992.

[GIBBON, D. *et al*. 2000] Gibbon, D., I. Mertins, and R. K. Moore (Eds.), *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Kluwer Academic Publishers, 2000.

[GRIOL, D. *et al*. 2008] Griol, D., L. F. Hurtado, E. Segarra, and E. Sanchis. *A statistical Approach to Spoken Dialog Systems Design and Evaluation*, Speech Communication, 50(8-9), 666-682, 2008.

[GRIOL, D. *et al*. 2011] Griol, D., J. Carbó, and J. M. Molina, *Agent Simulation to Develop Interactive and User-Centered Conversational Agents*, Advances in Intelligent and Soft Computing, 91, 69-76, 2011.

[GRIOL, D. *et al*. 2012] Griol, D., J. Molina, and Z. Callejas. *Bringing together commercial and academic perspectives for the development of intelligent AmI interfaces*, Journal of Ambient Intelligence and Smart Environments, 4(3), 83-207, 2012.

[HASEEL, L. *et al*. 2005] Haseel L, Hagen E. *Adaptation of an automotive dialogue system to users' expertise*. In: Proc. of 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech), Lisbon, Portugal, 222-226, 2005.

[HEIM, J. *et al*. 2005] Heim, J., Nilsson, E. G., Skjetne, J. H. *User Profiles for Adapting Speech Support in the Opera Web Browser to Disabled Users. LNCS*, 4397, 154-172, 2007.

[HEINROTH, T. *et al*. 2012] Heinroth, T. and W. Minker, *Introducing Spoken Dialogue Systems into Intelligent Environments*. Springer, 2012.

[JOKINEN, K., 2003] Jokinen, K. *Natural interaction in spoken dialogue systems*. In: Proc. of the Workshop Ontologies and Multilinguality in User Interfaces. Crete, Greece, 730-734, 2003.

[LALANNE, D. *et al*. 2009] Lalanne, D., L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, and J. Ladry. *Fusion engines for multimodal input: a survey*, in Proc. of ICMI-MLMI'09, 153-160, 2009.

[LANGNER, B. *et al*. 2005] Langner, B., Black, A. *Using speech in noise to improve understandability for elderly listeners*. In: Proc. of ASRU'05. San Juan, Puerto Rico, 392-396, 2005.

[LEMON, O. *et al*. 2012] Lemon, O. and O. Pietquin (Eds.), *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Computational Learning for Conversational Interfaces. Springer, 2012.

[LECH, T. *et al*. 2005] Lech, T. and L. W. M. Wienhofen, AmbieAgents: *A Scalable Infrastructure for Mobile and Context-Aware Information Services*. In: Proc. of AAMAS'05, 625-631, 2005.

[LEVIN, E. *et al*. 2006] Levin E, Levin A. *Dialog design for user adaptation*. In: Proc. of the International Conference on Acoustics Speech Processing, Toulouse, France, 57-60, 2006.

[LIGGINS, M. *et al*. 2009] Liggins, M., Hall, D., and Llinas, J. *Handbook of Multisensor Data Fusion* (2nd Edition). Boca Ratón, Florida, USA: CRC Press, 2009.

[LO, B.P. *et al*. 2003] Lo, B.P. J. Sun, and S. A. Velastin. *Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems*, Acta Automatica Sinica, 29(3), 393-407, 2003.

[LÓPEZ-CÓZAR, R. *et al*. 2005] López-Cózar, R. and M. Araki. *Spoken, Multilingual and Multimodal Dialogue Systems*. John Wiley & Sons Publishers, 2005.

[LÓPEZ-CÓZAR, R. *et al*. 2008] López-Cózar, R., and Callejas, Z. *ASR post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information*. Computer Speech and Language, 50, 745-766, 2008.

[MARKOPOULOS, P. *et al*. 2005] Markopoulos P, de Ruyter B, Privender S, van Breemen A. *Case study: bringing social intelligence into home dialogue systems*. Interactions, 12(4), 37-44, 2005.

[MARTINOVSKI, B. *et al*. 2003] Martinovski, B., Traum, D. *Breakdown in human-machine interaction: the error is the clue*. In: Proc. of the ISCA Tutorial and Research Workshop on Error Handling in Dialogue Systems. Chateau d'Oex, Vaud, Switzerland, 11-16, 2003.

[McCARTHY, J., 1987] McCarthy, J. *Generality in Artificial Intelligence*. Communications of the ACM, 30(12), 1030-1035, 1987.

[MINKER, W., 1998] Minker, W. *Stochastic versus rule-based speech understanding for information retrieval*. Speech Communication 25(4), 223-247, 1998.

[MINKER, W., 1999] Minker, W. *Design considerations for knowledge source representations of a stochastically-based natural language understanding component*. Speech Communication, 28, 141-154, 1999.

[NAZARI, AA., 2005] Nazari AA. *A Generic UPnP Architecture for Ambient Intelligence Meeting Rooms and a Control Point allowing for Integrated 2D and 3D Interaction*. In: Proc. of Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services, Usages and Technologies, 207-212, 2005.

[NIGAY, L. *et al*. 1995] Nigay L, Coutaz J. *A generic platform for addressing the multimodal challenge*. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Denver, Colorado, US, 98-105, 1995.

[OSLAND, P. *et al*. 2006] Osland, P., B. Viken, F. Solsvik, G. Nygreen, J. Wedvik, and S. Myklbust, *Enabling Context-Aware Applications*, In: Proc. of ICIN'06, 1-6, 2006.

[PIERACCINI, R., 2012] Pieraccini, R. *The Voice in the Machine: Building Computers that Understand Speech*. The MIT Press, 2012.

Rabiner, L., Juang, B. *Fundamentals of Speech Recognition*. Prentice Hal, 1993.

[PRENDINGER, H. *et al*. 2003] Prendinger, H., Mayer, S., Mori, J., Ishizuka, M. *Persona effect revisited. Using bio-signals to measure and reflect the impact of character-based interfaces*. In: Proc. of IVA'03. Kloster Irsee, Germany, 283-291, 2003.

[RADFORD, L., 2003] Radford, L. *Gestures, Speech, and the Sprouting of Signs: A Semiotic-Cultural Approach to Students' Types of Generalization*.

Mathematical thinking and learning, 5 (1), 37-70, 2003.

[RAUX, A. *et al*. 2003] Raux, A., Langner, B., Black, A. W., Eskenazi, M. *LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives*. In: Proc. of Eurospeech'03, Geneva, Switzerland, pp. 753-756, 2003.

[SALOVEY, P. *et al*. 1990] Salovey, P., Mayer, J.D. *Emotional intelligence*. Imagination, Cognition, and Personality, 9, 185-211, 1990.

[SÁNCHEZ, A.M. *et al*. 2007] Sánchez, A.M., M. Patricio, J. García, and J. M. Molina. *Video tracking improvement using context-based information*. In: Proc. of 10th Int. Conference on Information Fusion, 1-7, 2007.

[SCHATZMANN, J. *et al*. 2006] Schatzmann, J., K. Weilhammer, M. Stuttle, and S. Young. *A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies*, Knowledge Engineering Review, 21(2), 97–126, 2006.

[SCHULLER, S. *et al*. 2011] Schuller, B., Batliner, A., Steidl, S., Seppi, D. *Recognising Realistic Emotions and Affect in Speech: State Of The Art and Lessons Learnt from The First Challenge*. Speech Communication, vol.53(9-10), 1062-1087, 2011.

[SENEFF, S. *et al*. 2007] Seneff, S., M. Adler, J. Glass, B. Sherry, T. Hazen, C.Wang, and T.Wu. *Exploiting Context Information in Spoken Dialogue Interaction with Mobile Devices*. In: Proc. of IMUx'07, 1-11, 2007.

[STRAUSS, P. *et al*. 2010] Strauss, P. and W. Minker. *Proactive Spoken Dialogue Interaction in Multi-Party Environments*. Springer, 2010.

[TRAUM, D. *et al*. 2003] Traum, D., Larsson, S. *Current and New Directions in Discourse and Dialogue*, chap. The Information State Approach to Dialogue Management, pp. 325–354. Kluwer Academic Publishers, 2003.

[TSILFIDIS, A. *et al*. 2013] Tsilfidis, A., Mporas, I., Mourjopoulos, J., and Fakotakis, N. *Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing*. Computer Speech & Language, 27, 380-395, 2013.

[WAHLSTER, W., 2003] Wahlster, W. *Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture, and Facial Expression*. In: Proc. of the 26th German Conference on Artificial Intelligence, 1-18, 2003.

[WEBER, M.E. *et al*. 1994] Weber, M.E. and M. L. Stone. *Low altitude wind shear detection using airport surveillance radars*. In: Record of IEEE Radar Conference, 52-57, 1994.

[WILLIAMS, J., *et al*. 2007] Williams, J., Young, S. *Scaling POMDPs for Spoken Dialog Management*. IEEE Audio, Speech and Language Processing 15(8), 2116-2129, 2007.

[WOOLDRIDGE, M. *et al*. 1995] Wooldridge, M. and N. R. Jennings. *Surveillance and Canadian maritime domestic security*. The Knowledge Engineering Review, 10(2), 115-152, 1995.

[WU, L. *et al*. 2002] Wu, L., S. L. Oviatt, and P. R. Cohen. *From members to teams to committee-a robust approach to gestural and multimodal recognition*. IEEE Transactions on Neural Networks, 13(4), 972-982, 2002.

[WU, W.-L. *et al*. 2010] Wu, W.-L., Lu, R.-Z., Duan, J.-Y., Liu, H., Gao, F., and Chen, Y.-Q. *Spoken language understanding using weakly supervised learning. Computer Speech & Language*, *24*, 358-382, 2010.

[YOUNG, S., 2002] Young, S. *The Statistical Approach to the Design of Spoken Dialogue Systems*. Tech. rep., Cambridge University Engineering Department (UK), 2002.